Experiment no.1

AIM: - Data Collection- YouTube, connect to and capture social media

data for business (scraping, crawling, parsing).

Theory: In the digital era, businesses rely on social media analytics to understand audience behavior, trends, and engagement. YouTube, being one of the largest video-sharing platforms, provides valuable data, including video metadata, comments, views, likes, and subscriber counts.

Concepts:

1. **Web Scraping**
   - Extracts structured data from web pages using automated scripts.
   - Tools: BeautifulSoup, Selenium, Scrapy.
2. **Web Crawling**
   - Systematically browses the web to collect data from multiple pages.
   - Tools: Scrapy, Selenium, requests.
3. **Parsing**
   - Processes raw data into structured information.
   - JSON, XML, and HTML parsing techniques are used.

YouTube Data Collection Methods:

1. **Using YouTube Data API**
   - Official method, provides structured data but has rate limits.
   - Requires API key and authentication.
2. **Web Scraping YouTube**
   - Extracts data directly from YouTube’s web interface.
   - Helps bypass API limits but may face challenges due to dynamic content loading.

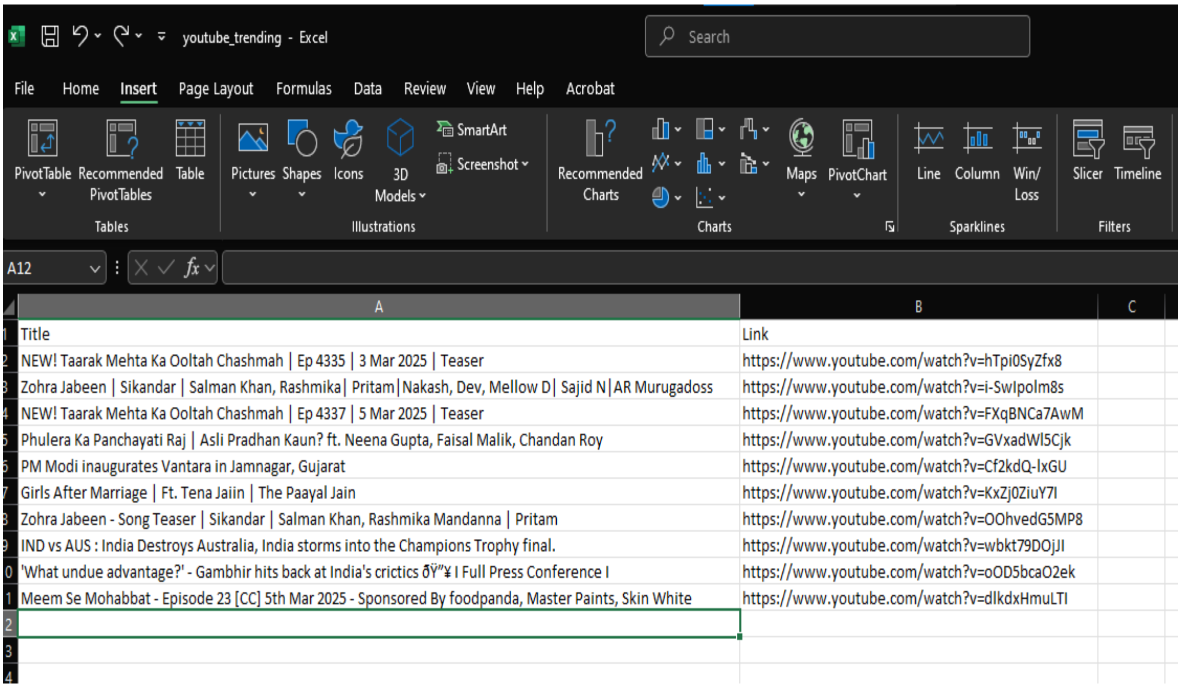
Applications in Business:

- Sentiment analysis from YouTube comments.

- Competitor analysis by tracking engagement on videos.
- Trend analysis using video views and search patterns.
- Influencer marketing insights.

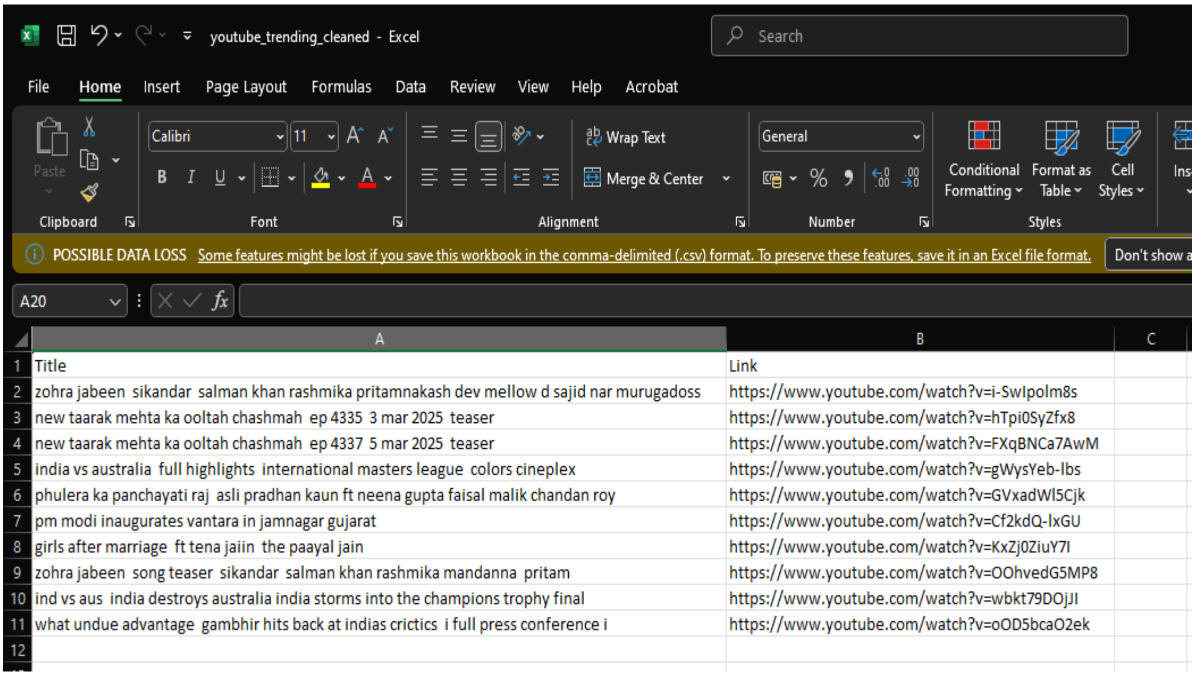
Code: -

```
from selenium import webdriver
from selenium.webdriver.chrome.service import Service
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.common.by import By
import time
import pandas as pd

# ◈ Automatically install the correct ChromeDriver version
service = Service(ChromeDriverManager().install())
driver = webdriver.Chrome(service=service)

# Open YouTube Trending Page
driver.get("https://www.youtube.com/feed/trending")
time.sleep(5)  # Wait for page to load

# Scrape video titles and links
videos = driver.find_elements(By.XPATH, '//a[@id="video-title"]')

video_data = []
for video in videos[:10]:  # Get top 10 videos
    title = video.text
    link = video.get_attribute("href")
    video_data.append({"Title": title, "Link": link})

# Save data to CSV
df = pd.DataFrame(video_data)
df.to_csv("youtube_trending.csv", index=False, encoding="utf-8")
print("✅ Data saved to youtube_trending.csv")

# Close browser
driver.quit()

# 🔹 Your existing scraping code (unchanged)
# (Scrapes data from YouTube and saves to youtube_trending.csv)

import pandas as pd
# Load CSV file
df = pd.read_csv("youtube_trending.csv")

# 🔹 Data Cleaning Steps
df = df.drop_duplicates()  # Remove duplicate videos
df = df.dropna(subset=['Title', 'Link'])  # Remove rows with missing values
df['Title'] = df['Title'].str.replace(r'[^\w\s]', '', regex=True)  # Remove special characters
df['Title'] = df['Title'].str.lower()  # Convert titles to lowercase

# Save cleaned data
df.to_csv("youtube_trending_cleaned.csv", index=False, encoding="utf-8")
print("✅ Data cleaned and saved to youtube_trending_cleaned.csv")
```



	A	B
1	Title	Link
2	NEW! Taarak Mehta Ka Ooltah Chashmah | Ep 4335 | 3 Mar 2025 | Teaser	https://www.youtube.com/watch?v=hTpi0SyZfx8
3	Zohra Jabeen | Sikandar | Salman Khan, Rashmika| Pritam|Nakash, Dev, Mellow D| Sajid N|AR Murugadoss	https://www.youtube.com/watch?v=i-SwIpolm8s
4	NEW! Taarak Mehta Ka Ooltah Chashmah | Ep 4337 | 5 Mar 2025 | Teaser	https://www.youtube.com/watch?v=FXqBNCa7AwM
5	Phulera Ka Panchayati Raj | Asli Pradhan Kaun? ft. Neena Gupta, Faisal Malik, Chandan Roy	https://www.youtube.com/watch?v=GVxadWl5Cjk
6	PM Modi inaugurates Vantara in Jamnagar, Gujarat	https://www.youtube.com/watch?v=Cf2kdQ-lxGU
7	Girls After Marriage | Ft. Tena Jaiin | The Paayal Jain	https://www.youtube.com/watch?v=KxZj0ZiuY7I
8	Zohra Jabeen - Song Teaser | Sikandar | Salman Khan, Rashmika Mandanna | Pritam	https://www.youtube.com/watch?v=OOhvedG5MP8
9	IND vs AUS : India Destroys Australia, India storms into the Champions Trophy final.	https://www.youtube.com/watch?v=wbkt79DOjJI
10	'What undue advantage?' - Gambhir hits back at India's crictics ðŸ”¥ I Full Press Conference I	https://www.youtube.com/watch?v=oOD5bcaO2ek
11	Meem Se Mohabbat - Episode 23 [CC] 5th Mar 2025 - Sponsored By foodpanda, Master Paints, Skin White	https://www.youtube.com/watch?v=dlkdxHmuLTI
12		
13		

Youtube_trending csv output



POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

	A	B
1	Title	Link
2	zohra jabeen sikandar salman khan rashmika pritamnakash dev mellow d sajid nar murugadoss	https://www.youtube.com/watch?v=i-SwIpolm8s
3	new taarak mehta ka ooltah chashmah ep 4335 3 mar 2025 teaser	https://www.youtube.com/watch?v=hTpi0SyZfx8
4	new taarak mehta ka ooltah chashmah ep 4337 5 mar 2025 teaser	https://www.youtube.com/watch?v=FXqBNCa7AwM
5	india vs australia full highlights international masters league colors cineplex	https://www.youtube.com/watch?v=gWysYeb-lbs
6	phulera ka panchayati raj asli pradhan kaun ft neena gupta faisal malik chandan roy	https://www.youtube.com/watch?v=GVxadWl5Cjk
7	pm modi inaugurates vantara in jamnagar gujarat	https://www.youtube.com/watch?v=Cf2kdQ-lxGU
8	girls after marriage ft tena jaiin the paayal jain	https://www.youtube.com/watch?v=KxZj0ZiuY7I
9	zohra jabeen song teaser sikandar salman khan rashmika mandanna pritam	https://www.youtube.com/watch?v=OOhvedG5MP8
10	ind vs aus india destroys australia india storms into the champions trophy final	https://www.youtube.com/watch?v=wbkt79DOjJI
11	what undue advantage gambhir hits back at indias crictics i full press conference i	https://www.youtube.com/watch?v=oOD5bcaO2ek
12		

Youtube_trending_cleaned csv output